High-dimensional stochastic optimization with the generalized Dantzig estimator

Karim Lounici

November 17, 2008

Abstract

We propose a generalized version of the Dantzig selector. We show that it satisfies sparsity oracle inequalities in prediction and estimation. We consider then the particular case of high-dimensional linear regression model selection with the Huber loss function. In this case we derive the sup-norm convergence rate and the sign concentration property of the Dantzig estimators under a mutual coherence assumption on the dictionary.

Key words: Dantzig, Sparsity, Prediction, Estimation, Sign consistency.
2000 Mathematics Subject Classification Primary: 62G25, 62G05; Secondary: 62J05, 62J12.

1 Introduction

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ be a measurable space. We observe a set of n i.i.d. random pairs $Z_i = (X_i, Y_i)$, $i = 1, \dots, n$ where $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$. Denote by P the joint distribution of (X_i, Y_i) on $\mathcal{X} \times \mathcal{Y}$, and by P^X the marginal distribution of X_i. Let $Z = (X, Y)$ be a random pair in $\mathcal{Z}$ distributed according to P. For any real-valued function g on $\mathcal{X}$, define $||g||_\infty = \operatorname{ess\,sup}_{x \in \mathcal{X}} |g(x)|$, $\|g\| = \left(\int_{\mathcal{X}} g(x)^2 P^X(dx)\right)^{1/2}$ and $||g||_n = \left(\frac{1}{n}\sum_{i=1}^n g(X_i)^2\right)^{1/2}$. Let $\mathcal{D} = \{f_1, \dots, f_M\}$ be a set of real-valued functions on $\mathcal{X}$ called the dictionary where $M \geqslant 2$. We assume that the functions of the dictionary are normalized, so that $\|f_j\| = 1$ for all $j = 1, \dots, M$. We also assume that $||f_j||_\infty \leqslant L$ for some $L > 0$. For any $\theta \in \mathbb{R}^M$, define $f_\theta = \sum_{j=1}^M \theta_j f_j$ and $J(\theta) = \{j \ : \ \theta_j \neq 0\}$. Let $M(\theta) = |J(\theta)|$ be the cardinality of $J(\theta)$ and $\vec{\operatorname{sign}}(\theta) = (\operatorname{sign}(\theta_1), \dots, \operatorname{sign}(\theta_M))^T$ where

$$\operatorname{sign}(t) = \begin{cases} 1 & \text{if } t > 0, \\ 0 & \text{if } t = 0, \\ -1 & \text{if } t < 0. \end{cases}$$

For any vector $\theta \in \mathbb{R}^M$ and any subset J of $\{1, \dots, M\}$, we denote by θ_J the vector in $\mathbb{R}^M$ which has the same coordinates as θ on J and zero coordinates on

the complement J^c of J. For any integers $1 \leqslant d, p < \infty$ and $w = (w_1, \ldots, w_d) \in \mathbb{R}^d$, the l_p norm of the vector w is denoted by $|w|_p \triangleq \left(\sum_{j=1}^d |w_j|^p\right)^{1/p}$, and $|w|_\infty \triangleq \max_{1 \leqslant j \leqslant d} |w_j|$.

Consider a function $\gamma : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}^+$ such that for any y in $\mathcal{Y}$ and u, u' in $\mathbb{R}$ we have

$$|\gamma(y, u) - \gamma(y, u')| \leqslant |u - u'|.$$

We assume furthermore that $\gamma(y, \cdot)$ is convex and differentiable for any $y \in \mathcal{Y}$. We assume that for any $y \in \mathcal{Y}$ the derivative $\partial_u \gamma(y, \cdot)$ is absolutely continuous. Then $\partial_u \gamma(y, \cdot)$ admits a derivative almost everywhere which we denote by $\partial_u^2 \gamma(y, \cdot)$. Consider the loss function $Q : \mathcal{Z} \times \mathbb{R}^M \to \mathbb{R}^+$ defined by

$$Q(z, \theta) = \gamma(y, f_\theta(x)). \tag{1}$$

The expected and empirical risk measures at point θ in $\mathbb{R}^M$ are defined respectively by

$$R(\theta) \triangleq \mathbb{E}\left(Q(Z, \theta)\right),$$

where $\mathbb{E}$ is the expectation sign, and

$$\hat{R}_n(\theta) \triangleq \frac{1}{n} \sum_{i=1}^n Q(Z_i, \theta).$$

Define the target vector as a minimizer of $R(\cdot)$ over $\mathbb{R}^M$:

$$\theta^* \triangleq \arg \min_{\theta \in \mathbb{R}^M} R(\theta).$$

Note that the target vector is not necessarily unique. From now on, we assume that there exists a s-sparse solution θ^*, i.e., a solution with $M(\theta^*) \leqslant s$, and that this sparse solution is unique. We will see that this is indeed the case under the coherence condition on the dictionary (cf. Section 3 below).

Define the excess risk of the vector θ by

$$\mathcal{E}(\theta) = R(\theta) - R(\theta^*),$$

and its empirical version by

$$\mathcal{E}_n(\theta) = R_n(\theta) - R_n(\theta^*).$$

Our goal is to derive sparsity oracle inequalities for the excess risk and for the risk of θ^* in the l_1 norm and in the sup-norm.

We consider the following minimization problem:

$$\min_{\theta \in \Theta} |\theta|_1 \ \text{ subject to } \ \left|\nabla \hat{R}_n(\theta)\right|_\infty \leqslant r, \tag{2}$$

where $\nabla \hat{R}_n \triangleq (\partial_{\theta_1} \hat{R}_n, \ldots, \partial_{\theta_M} \hat{R}_n)^T$, $r > 0$ is a tuning parameter defined later and Θ is a convex subset of $\mathbb{R}^M$ specified later. Solutions of (2), if they exist,

will be taken as estimators of θ^*. Note that we will prove in Lemma 3 that under Assumption 2 the set $\{\theta \in \Theta : \left|\nabla \hat{R}_n(\theta)\right|_\infty \leqslant r\}$ is non-empty with probability close to one. Note also that in the applications considered in Section 3, the constraint $|\nabla \hat{R}_n(\theta)|_\infty \leqslant r$ can be defined as a system of inequalities involving convex functions. Thus, solutions to (2) exist and can be efficiently computed via convex optimization. In particular, for the regression model with the Huber loss, the gradient $\nabla \hat{R}_n(\theta)$ is piecewise linear so that (2) reduces in this case to a standard linear programming problem. Denote by $\hat{\Theta}$ the set of all solutions of (2). For the reasons above, we assume from now on that $\hat{\Theta} \neq \emptyset$ with probability close to one.

The definition of our estimator (2) can be motivated as follows. Since the loss function $Q(z, \cdot)$ is convex and differentiable for any fixed $z \in \mathcal{Z}$, the expected risk R is also a convex function of θ and it is differentiable under mild conditions. Thus, minimizing R is equivalent to finding the zeros of ∇R. The quantity $\nabla \hat{R}_n(\theta)$ is the empirical version of $\nabla R(\theta)$. We choose the constant r such that the vector θ^* satisfies the constraint $|\nabla \hat{R}_n(\theta^*)| \leqslant r$ with probability close to 1. Then among all the vectors satisfying this constraint, we choose those with minimum l_1 norm. Note that if we consider the linear regression problem with the quadratic loss, we recognize in (2) the Dantzig minimization problem of Candes and Tao [7]. From now on, we will call (2) the generalized Dantzig minimization problem.

Bickel et al. [1], Candes and Tao [7] and Koltchinskii [12] proved that the Dantzig estimator performs well in high-dimensional regression problems with the quadratic loss. In particular they proved sparsity oracle inequalities on the excess risk and the estimation of θ^* for the l_p norm with $1 \leqslant p \leqslant 2$.

The problem (2) is closely related to the minimization problem:

$$\min_{\theta \in \Theta} \hat{R}_n(\theta) + r|\theta|_1, \tag{3}$$

which is a generalized version of the Lasso. For the Lasso estimator, Bunea et al [5] proved similar results in high-dimensional regression problems with the quadratic loss under a mutual coherence assumption [8] and Bickel et al [1] under a weaker Restricted Eigenvalue assumption. Koltchinskii [11] derived similar results for the Lasso in the context of high-dimensional regresssion with twice differentiable Lipschtiz continuous loss functions under a restricted isometry assumption. Van de Geer [22, 23] obtained similar results for the Lasso in the context of generalized linear models with Lipschtiz continuous loss functions. Lounici [15] derived sup-norm convergence rates and sign consistency of the Lasso and Dantzig estimators in a high-dimensional linear regression model with the quadratic loss under a mutual coherence assumption.

The paper is organized as follows. In Section 2 we derive sparsity oracle inequalities for the excess risk and for estimation of θ^* for the generalized Dantzig estimators defined by (2) in a stochastic optimization framework. In section 3 we apply the results of Section 2 to the linear regression model with the Huber loss and to the logistic regression model. In Section 4 we prove the variable selection consistency with rates under a mutual coherence assumption for the

linear regression model with the Huber loss. In section 5 we show a sign concentration property of the thresholded generalized Dantzig estimators for the linear regression model with the Huber loss.

2 Sparsity oracle inequalities for prediction and estimation with the l_1 norm

We need an assumption on the dictionary to derive prediction and estimation results for the generalized Dantzig estimators. We first state the Restricted Eigenvalue assumption [1].

Assumption 1.

$$\zeta(s) \stackrel{\triangle}{=} \min_{J_0 \subset \{1,\ldots,M\}:|J_0| \leqslant s} \quad \min_{\Delta \neq 0:|\Delta_{J_0^c}|_1 \leqslant |\Delta_{J_0}|_1} \frac{||f_\Delta||}{|\Delta_{J_0}|_2} > 0.$$

It implies an "equivalence" between the two norms $|\Delta|_2$ and $\|f_\Delta\|$ on the subset $\{\Delta \neq 0 : |\Delta_{J(\Delta)^c}|_1 \leqslant |\Delta_{J(\Delta)}|_1\}$ of $\mathbb{R}^M$.

We need the following assumption on $\|f_{\theta^*}\|_\infty$.

Assumption 2. *There exists a constant $K > 0$ such that $\|f_{\theta^*}\|_\infty \leqslant K$.*

From now on we take for Θ the set

$$\Theta = \{\theta \in \mathbb{R}^M \, : \, \|f_\theta\|_\infty \leqslant K\}.$$

The following assumption is a version of the margin condition (cf. [21]). It links the excess risk to the functional norm $\|\cdot\|$.

Assumption 3. *For any $\theta \in \Theta$ there exits a constant $c > 0$ depending possibly on K such that*

$$\|f_\theta - f_{\theta^*}\| \leqslant c(R(\theta) - R(\theta^*))^{1/\kappa},$$

where $1 < \kappa \leqslant 2$.

We will prove in Section 2.1 below that this condition is always satisfied with the constant $\kappa = 2$ for the regression model with Huber loss and for the logistic regression model. We also need the following technical assumption.

Assumption 4. *The constants K and L satisfy*

$$1 \leqslant K, L \leqslant \sqrt{\frac{n}{\log M}}.$$

Define the quantity

$$\tilde{r} = 4\sqrt{2}L\frac{\log M}{n} + 2\sqrt{6}\sqrt{\frac{\log M}{n}}. \tag{4}$$

We assume from now on that $\tilde{r} \leqslant 1$.

The main results of this section are the following sparsity oracle inequalities for the excess risk and for estimation of θ^* in the l_1 norm. Define

$$r = 6\|\partial_u \gamma\|_\infty \tilde{r}. \tag{5}$$

Theorem 1. *Let Assumptions 1 - 4 be satisfied. Take r as in (5). Assume that $M(\theta^*) \leqslant s$. Then, with probability at least $1 - M^{-1} - M^{-K} - 3M^{-2K} \log \frac{n}{\log M}$, we have*

$$\sup_{\hat{\theta} \in \hat{\Theta}} \mathcal{E}(\hat{\theta}) \leqslant \left(\frac{2(1+2K)cr\sqrt{s}}{\zeta(s)} \right)^{\frac{\kappa}{\kappa-1}} + 12\|\partial_u \gamma\|_\infty \frac{\kappa}{\kappa - 1} \tilde{r}^2, \tag{6}$$

and

$$\sup_{\hat{\theta} \in \hat{\Theta}} |\hat{\theta} - \theta^*|_1 \leqslant \left(\frac{2c\sqrt{s}}{\zeta(s)} \right)^{\frac{\kappa}{\kappa-1}} ((1+2K)r)^{\frac{1}{\kappa-1}} + \frac{2K}{(\kappa-1)(1+2K)} \tilde{r}. \tag{7}$$

Note that the regularization parameter r does not depend on the variance of the noise if we consider the regression model with non-quadratic loss. In this case, the use of Lipschtiz losses enables us to treat cases where the noise variable does not admit a finite second moment, e.g., the Cauchy distribution. The price to pay is that we need to assume that $\|f_{\theta^*}\|_\infty \leqslant K$ with known K.

Proof. For any $\hat{\theta} \in \hat{\Theta}$ define $\Delta = \hat{\theta} - \theta^*$. We have

$$\begin{aligned} \mathcal{E}(\hat{\theta}) &\leqslant \mathcal{E}_n(\hat{\theta}) + \mathcal{E}(\hat{\theta}) - \mathcal{E}_n(\hat{\theta}) \\ &= \mathcal{E}_n(\hat{\theta}) + \frac{\mathcal{E}(\hat{\theta}) - \mathcal{E}_n(\hat{\theta})}{|\Delta|_1 + \tilde{r}} (|\Delta|_1 + \tilde{r}) \\ &\leqslant \mathcal{E}_n(\hat{\theta}) + \sup_{\theta \in \Theta : \theta \neq \theta^*} \left(\frac{\mathcal{E}(\theta) - \mathcal{E}_n(\theta)}{|\theta - \theta^*|_1 + \tilde{r}} \right) (|\Delta|_1 + \tilde{r}). \end{aligned} \tag{8}$$

By Lemma 1 it holds on an event $\mathcal{A}_1$ of probability at least $1 - M^{-K} - 3M^{-2K} \log \frac{n}{\log M}$ that

$$\sup_{\theta \in \Theta : \theta \neq \theta^*} \frac{\mathcal{E}(\theta) - \mathcal{E}_n(\theta)}{|\theta - \theta^*|_1 + \tilde{r}} \leqslant 2Kr. \tag{9}$$

For any $\hat{\theta} \in \hat{\Theta}$, we have by definition of the Dantzig estimator that $|\hat{\theta}|_1 \leqslant |\theta^*|_1$. Thus

$$|\Delta_{J(\theta^*)^c}|_1 = \sum_{j \in J(\theta^*)^c} |\hat{\theta}_j| \leqslant \sum_{j \in J(\theta^*)} |\theta^*_j| - |\hat{\theta}_j| \leqslant |\Delta_{J(\theta^*)}|_1. \tag{10}$$

Define the function $g : t \to R_n(\theta^* + t\Delta)$. Clearly g is convex and differentiable on $[0, 1]$. Thus, the function g' is nondecreasing on $[0, 1]$ with derivative

$g'(t) = \nabla R_n(\theta^* + t\Delta)^T \Delta$. The constraint $\left|\nabla \hat{R}_n(\theta)\right|_\infty \leqslant r$ in (2) and Lemma 3 yield, on an event $\mathcal{A}_2$ of probability at least $1 - M^{-1}$,

$$\begin{aligned} \mathcal{E}_n(\hat{\theta}) &= R_n(\hat{\theta}) - R_n(\theta^*) \\ &= \int_0^1 \nabla R_n(\theta^* + t\Delta)^T \Delta dt \\ &\leqslant r|\Delta|_1, \end{aligned} \tag{11}$$

for some numerical constant $C > 0$.

Combining (8)-(11) yields that on the event $\mathcal{A}_1 \cap \mathcal{A}_2$

$$\mathcal{E}(\hat{\theta}) \leqslant (2 + 4K) r |\Delta_{J(\theta^*)}|_1 + 12 \|\partial_u \gamma\|_\infty K \tilde{r}^2. \tag{12}$$

Next,

$$\begin{aligned} 2(1+2K)r|\Delta_{J(\theta^*)}|_1 &\leqslant 2(1+2K)r\sqrt{s}|\Delta_{J(\theta^*)}|_2 \\ &\leqslant \frac{2(1+2K)cr\sqrt{s}}{\zeta(s)} \frac{\|f_\Delta\|}{c} \\ &\leqslant \frac{1}{\kappa'} \left(\frac{2cr\sqrt{s}}{\zeta(s)} \right)^{\kappa'} + \frac{1}{\kappa} \left(\frac{\|f_\Delta\|}{c} \right)^{\kappa} \\ &\leqslant \frac{1}{\kappa'} \left(\frac{2(1+2K)cr\sqrt{s}}{\zeta(s)} \right)^{\kappa'} + \frac{1}{\kappa} \mathcal{E}(\hat{\theta}^D), \end{aligned} \tag{13}$$

where we have used the Cauchy-Schwarz inequality in the first line, the inequality $xy \leqslant |x|^\kappa/\kappa + |y|^{\kappa'}/\kappa'$ that holds for any x, y in $\mathbb{R}$ and for any κ, κ' in $(1, \infty)$ such that $1/\kappa + 1/\kappa' = 1$ in the third line, and Assumption 2 in the last line. Combining (12) and (13) and the fact that $\tilde{r} \leqslant 1$ yields the first inequality. The second inequality is a consequence of (6) and (13). □

We state and prove below intermediate results used in the proof of Theorem 1.

Lemma 1. *Let Assumptions 2 and 4 be satisfied. Then, with probability at least* $1 - M^{-K} - 3M^{-2K} \log \frac{n}{\log M}$, *we have*

$$\sup_{\theta \in \Theta} \frac{|\mathcal{E}(\theta) - \mathcal{E}_n(\theta)|}{|\theta - \theta^*|_1 + \tilde{r}} \leqslant 2Kr, \tag{14}$$

where r is defined in Theorem 1.

Proof. For any $A > 0$, define the random variable

$$T_A = \sup_{\theta \in \Theta : |\theta - \theta^*|_1 \leqslant A} |\mathcal{E}_n(\theta) - \mathcal{E}(\theta)|.$$

For any θ in Θ and (x, y) in $\mathcal{Z}$ we have

$$|\gamma(y, f_\theta(x)) - \gamma(y, f_{\theta^*}(x))| \leqslant \|\partial_u \gamma\|_\infty \left(L|\theta - \theta^*|_1 \wedge 2K \right),$$

and

$$\mathbb{E}\left(|\gamma(Y, f_\theta(X)) - \gamma(Y, f_{\theta^*}(X))|^2\right) \leqslant \|\partial_u \gamma\|_\infty^2 \left(|\theta - \theta^*|_1^2 \wedge 2K^2\right).$$

Assumption 3 and Bousquet's concentration inequality (cf. Theorem 4 in Section 6 below) with $x = (A \vee 2K)\log M$, $c = 2\|\partial_u\gamma\|_\infty(AL \wedge 2K)$ and $\sigma = \sqrt{2}\|\partial_u\gamma\|_\infty(A \wedge \sqrt{2}K)$ yield

$$\mathbb{P}\left(T_A \geqslant \mathbb{E}(T_A) + 2AK\|\partial_u\gamma\|_\infty \tilde{r}\right) \leqslant M^{-(2K)\vee A}.$$

We study now the quantity $\mathbb{E}(T_A)$. By standard symmetrization and contraction arguments (cf. Theorems 5 and 6 in Section 6) we obtain

$$\mathbb{E}(T_A) \quad \leqslant \quad 4\|\partial_u\gamma\|_\infty \mathbb{E}\left(\sup_{\theta\in\Theta\,:\,|\theta-\theta^*|_1\leqslant A}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_{\theta-\theta^*}(X_i)\right|\right).$$

Then, observe that the mapping $u \to \frac{1}{n}\sum_{i=1}^n \epsilon_i f_u(X_i)$ is linear, thus its supremum on a simplex is attained at one of its vertices. This yields

$$\mathbb{E}(T_A) \leqslant 4\|\partial_u\gamma\|_\infty A \mathbb{E}\left(\max_{1\leqslant j\leqslant M}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_j(X_i)\right|\right).$$

Combining Assumption 4 and Lemma 2 we obtain

$$\mathbb{E}(T_A) \leqslant 4\|\partial_u\gamma\|_\infty A\tilde{r}.$$

Thus

$$\mathbb{P}\left(T_A \geqslant 6AK\|\partial_u\gamma\|_\infty\tilde{r}\right) \leqslant M^{-(2K)\vee A}. \tag{15}$$

Define the following subsets of Θ

$$\begin{aligned}
\Theta(I) &= \{\theta\in\Theta\,:\,|\theta-\theta^*|_1 \leqslant \tilde{r}\},\\
\Theta(II) &= \{\theta\in\Theta\,:\,\tilde{r} < |\theta-\theta^*|_1 \leqslant 2K\},\\
\Theta(III) &= \{\theta\in\Theta\,:\,|\theta-\theta^*|_1 > 2K\}.
\end{aligned}$$

For any $t > 0$ define the probabilities

$$\begin{aligned}
P_I &= \mathbb{P}\left(\sup_{\theta\in\Theta(I)}\frac{|\mathcal{E}(\theta)-\mathcal{E}_n(\theta)|}{|\theta-\theta^*|_1+\tilde{r}} \geqslant t\right)\\
P_{II} &= \mathbb{P}\left(\sup_{\theta\in\Theta(II)}\frac{|\mathcal{E}(\theta)-\mathcal{E}_n(\theta)|}{|\theta-\theta^*|_1+\tilde{r}} \geqslant t\right)\\
P_{III} &= \mathbb{P}\left(\sup_{\theta\in\Theta(III)}\frac{|\mathcal{E}(\theta)-\mathcal{E}_n(\theta)|}{|\theta-\theta^*|_1+\tilde{r}} \geqslant t\right)
\end{aligned}$$

For any $t > 0$ we have

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\frac{|\mathcal{E}(\theta)-\mathcal{E}_n(\theta)}{|\theta-\theta^*|_1+\tilde{r}} \geqslant t\right) \leqslant P_I + P_{II} + P_{III}.$$

Now, we bound from above the three probabilities on the right hand side of the above expression. Take $t = 12\|\partial_u \gamma\|_\infty K\tilde{r}$. Applying (15) to P_I yields that

$$P_I \leqslant \mathbb{P}\left(T_{\tilde{r}} \geqslant 6\|\partial_u \gamma\|_\infty K\tilde{r}^2\right) \leqslant M^{-2K},$$

since we have $\tilde{r} \leqslant K$ by Assumption 4.

Consider now P_{II}. We have

$$\Theta(II) \subset \bigcup_{j=0}^{j_0} \left\{\theta \in \Theta \,:\, A_{j+1} \leqslant |\theta - \theta^*|_1 \leqslant A_j\right\},$$

where $A_j = 2^{1-j}K$, $j = 0, \ldots, j_0$ and j_0 is chosen such that $2^{1-j_0}K > \tilde{r}$ and $2^{-j_0}K \leqslant \tilde{r}$. Thus

$$\begin{aligned}
P_{II} &\leqslant \sum_{j=0}^{j_0} \mathbb{P}\left(T_{A_j} \geqslant 12\|\partial_u \gamma\|_\infty A_{j+1} K\tilde{r}\right) \\
&\leqslant \sum_{j=0}^{j_0} \mathbb{P}\left(T_{A_j} \geqslant 6\|\partial_u \gamma\|_\infty A_j K\tilde{r}\right) \\
&\leqslant (j_0+1)M^{-2K} \\
&\leqslant \left(3\left(\log\frac{n}{\log M}\right) - 1\right) M^{-2K}.
\end{aligned}$$

Consider finally P_{III}. We have

$$\Theta(III) \subset \bigcup_{j=0}^{\infty} \left\{\theta \in \Theta \,:\, \bar{A}_{j-1} < |\theta - \theta^*|_1 \leqslant \bar{A}_j\right\},$$

where $\bar{A}_j = 2^{1+j}K$, $j \geqslant 0$. Thus

$$\begin{aligned}
P_{III} &\leqslant \sum_{j=1}^{\infty} \mathbb{P}\left(T_{\bar{A}_j} \geqslant 12\|\partial_u \gamma\|_\infty \bar{A}_{j-1} K\tilde{r}\right) \\
&\leqslant \sum_{j=0}^{j_0} \mathbb{P}\left(T_{A_j} \geqslant 6\|\partial_u \gamma\|_\infty \bar{A}_j K\tilde{r}\right) \\
&\leqslant \sum_{j=1}^{\infty} M^{-\bar{A}_j} \\
&\leqslant M^{-K}.
\end{aligned}$$

□

We now study the quantity $\mathbb{E}\left(\max_{1\leqslant j\leqslant M} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_j(X_i)\right|\right)$. This is done in the next lemma.

Lemma 2. *We have*

$$\mathbb{E}\left(\max_{1\leqslant j\leqslant M}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f_j(X_i)\right|\right)\leqslant \tilde{r}, \tag{16}$$

where $\tilde{r}$ is defined in (4).

Proof. Define the random variables

$$U_j=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i f_j(X_i).$$

The Bernstein inequality yields, for any $j=1,\dots,M$ and $t>0$,

$$\mathbb{P}\left(|U_j|\geqslant t\right)\leqslant \exp\left(-\frac{t^2}{2(t\|f_j\|_\infty/(3\sqrt{n})+\|f_j\|^2)}\right). \tag{17}$$

Set $b_j=\|f_j\|_\infty/(3\sqrt{n})$. Define the random variables $T_j=U_j\mathbb{1}_{|Y_j|>\|f_j\|^2/b_j}$ and $T_j'=U_j\mathbb{1}_{|Y_j|\leqslant\|f_j\|^2/b_j}$. For all $t>0$ we have

$$\mathbb{P}\left(|T_j|>t\right)\leqslant 2\exp\left(-\frac{t}{4b_j}\right),\quad \mathbb{P}\left(|T_j'|>t\right)\leqslant 2\exp\left(-\frac{t^2}{4\|f_j\|^2}\right).$$

Define the function $h_\nu(x)=\exp(x^\nu)-1$, where $\nu>0$. This function is clearly convex for any $\nu>0$. We have

$$\mathbb{E}\left(h_1\left(\frac{|T_j|}{12b_j}\right)\right)=\int_0^\infty e^t\mathbb{P}(|T_j|>12b_jt)dt\leqslant 1,$$

where we have used Fubini's Theorem in the first equality. Since the function h_1 is convex and nonnegative, we have

$$\begin{aligned}
h_1\left(\mathbb{E}\left(\max_{1\leqslant j\leqslant M}\frac{|T_j|}{12b_j}\right)\right) &\leqslant \mathbb{E}\left(h_1\left(\max_{1\leqslant j\leqslant M}\frac{|T_j|}{12b_j}\right)\right)\\
&\leqslant \mathbb{E}\left(\sum_{j=1}^{M}h_1\left(\frac{|T_j|}{12b_j}\right)\right)\\
&\leqslant M,
\end{aligned}$$

where we have used the Jensen inequality. Since the function $h_1^{-1}(x)=\log(1+x)$ is increasing, we have

$$\begin{aligned}
\mathbb{E}\left(\max_{1\leqslant j\leqslant M}\frac{|T_j|}{12b_j}\right) &\leqslant \log(1+M)\\
\mathbb{E}\left(\max_{1\leqslant j\leqslant M}|T_j|\right) &\leqslant 4\frac{\log(1+M)}{\sqrt{n}}\max_{1\leqslant j\leqslant M}\|f_j\|_\infty.
\end{aligned} \tag{18}$$

Applying the same argument to the function h_2, we prove that

$$\mathbb{E}\left(\max_{1\leqslant j\leqslant M}|T_j'|\right)\leqslant 2\sqrt{3}\sqrt{\log(1+M)}\max_{1\leqslant j\leqslant M}\|f_j\|. \tag{19}$$

Combining (18) and (19) yields the result.

□

Lemma 3. *Let Assumptions 2 and 4 be satisfied. Then, with probability at least* $1-M^{-1}$*, we have*

$$|\nabla\hat{R}_n(\theta^*)|_\infty\leqslant r,$$

where r *is defined in Theorem 1.*

Proof. For any $1\leqslant j\leqslant M$ define

$$Z_j=\frac{1}{n}\sum_{i=1}^n\partial_u\gamma(Y_i,f_{\theta^*}(X_i))f_j(X_i).$$

Since the function $\theta\to\gamma(y,f_\theta(x))$ is differentiable w.r.t. θ and $|\partial_u\gamma(y,f_\theta(x))f_j(x)|\leqslant\|\partial_u\gamma\|_\infty L$ for any $(x,y)\in\mathcal{X}\times\mathcal{Y}$ and $\theta\in\mathbb{R}^M$, we have

$$\mathbb{E}(Z_j)=\frac{\partial R(\theta^*)}{\partial\theta_j}=0.$$

Next, similarly as in Lemmas 1 and 2, we prove that

$$\mathbb{E}(|\nabla\hat{R}_n(\theta^*)|_\infty)\leqslant 4\|\partial_u\gamma\|_\infty\tilde{r}.$$

Finally Bousquet's concentration inequality (cf. Theorem 4 in Section 6 below) yields that, with probability at least $1-M^{-1}$,

$$\begin{aligned}|\nabla\hat{R}_n(\theta^*)|_\infty&\leqslant\mathbb{E}(|\nabla\hat{R}_n(\theta^*)|_\infty)\\&\quad+\sqrt{2\frac{\log M}{n}\left(\|\partial_u\gamma\|_\infty^2+2\|\partial_u\gamma\|_\infty L\mathbb{E}(|\nabla\hat{R}_n(\theta^*)|_\infty)\right)}\\&\quad+\frac{\|\partial_u\gamma\|_\infty L\log M}{3n}\\&\leqslant 6\|\partial_u\gamma\|_\infty\tilde{r}.\end{aligned}$$

□

3 Examples

3.1 Robust regression with the Huber loss

We consider the linear regression model

$$Y=f_{\theta^*}(X)+W, \tag{20}$$

where $X \in \mathbb{R}^d$ is a random vector, $W \in \mathbb{R}$ is a random variable independent of X whose distribution is symmetric w.r.t. 0 and $\theta^* \in \mathbb{R}^M$ is the unknown vector of parameters. Consider the function

$$\phi(x) = \frac{1}{2}x^2 \mathbb{1}_{|x| \leqslant 2K+\alpha} + \left((2K+\alpha)|x| - \frac{(2K+\alpha)^2}{2}\right) \mathbb{1}_{|x|>2K+\alpha},$$

where $\alpha > 0$. The Huber loss function is defined by

$$Q(z, \theta) = \phi(y - f_\theta(x)), \tag{21}$$

where $z = (x, y) \in \mathbb{R}^d \times \mathbb{R}$ and $\theta \in \Theta$.

In the following lemma we prove that for this loss function Assumption 3 is satisfied with $\kappa = 2$ and $c = (2/\mathbb{P}(|W| \leqslant \alpha))^{1/2}$.

Lemma 4. *Let Q be defined by (21). Then for any $\theta \in \Theta$ we have*

$$\frac{\mathbb{P}(|W| \leqslant \alpha)}{2} \|f_\theta - f_{\theta^*}\|^2 \leqslant \mathcal{E}(\theta).$$

Proof. Set $\Delta = \theta - \theta^*$. Since ϕ' is absolutely continuous, we have for any $\theta \in \Theta$

$$\begin{aligned} Q(Z, \theta) - Q(Z, \theta^*) &= \phi'(W) f_{-\Delta}(X) \\ &\quad + \left[\int_0^1 \mathbb{1}_{|W + t f_{-\Delta}(X)| \leqslant 2K+\alpha}(1-t)dt\right] f_\Delta(X)^2 \\ &\geqslant \phi'(W) f_{-\Delta}(X) + \frac{1}{2}\mathbb{1}_{(|W| \leqslant \alpha)} f_\Delta(X)^2, \end{aligned}$$

since $\|f_\theta\|_\infty \leqslant K$ for any $\theta \in \Theta$. Taking the expectations we get

$$R(\theta) - R(\theta^*) \geqslant \frac{\mathbb{P}(|W| \leqslant \alpha)}{2} \|f_\Delta\|^2,$$

for any $\alpha > 0$ since ϕ' is odd and the distribution of W is symmetric w.r.t. 0. □

We have the following corollary of Theorem 1.

Corollary 1. *Let Assumptions 1, 2 and 4 be satisfied. If $M(\theta^*) \leqslant s$, then, with probability at least $1 - M^{-1} - M^{-K} - 3M^{-2K} \log \frac{n}{\log M}$, we have*

$$\sup_{\hat{\theta} \in \hat{\Theta}} \mathcal{E}(\hat{\theta}) \leqslant \frac{8(1+2K)^2}{\mathbb{P}(|W| \leqslant \alpha)\zeta(s)^2} sr^2 + \frac{2}{3}r^2,$$

and

$$\sup_{\hat{\theta} \in \hat{\Theta}} |\hat{\theta} - \theta^*|_1 \leqslant \frac{8(1+2K)}{\mathbb{P}(|W| \leqslant \alpha)\zeta(s)^2} sr + \frac{K}{3(1+2K)}r.$$

3.1.1 Logistic regression and similar models

We consider $Z = (X, Y) \in \mathcal{X} \times \{0, 1\}$ where $\mathcal{X}$ is a Borel subset of $\mathbb{R}^d$. The conditional probability $\mathbb{P}(Y = 1 \,|X = x) = \pi(x)$ is unknown where π is a function on $\mathcal{X}$ with values in $[0, 1]$. We assume that π is of the form

$$\pi(x) = \Phi'(f_{\theta^*}(x)), \tag{22}$$

where the function $\Phi : \mathbb{R} \to \mathbb{R}^*$ is convex, twice differentiable, of derivative Φ' with values in $[0, 1]$ and the vector $\theta^* \in \mathbb{R}^M$ is unknown. Consider, e.g., the logit loss function $\Phi(u) = \log(1 + e^u)$. We assume that Φ is known. Define the quantity

$$\tau(R) = \frac{1}{2} \inf_{|u| \leqslant R} \Phi^{(2)}(u), \tag{23}$$

for any $R \geqslant 0$. We want to estimate θ^* with the procedure (2) and the convex loss function

$$Q(z, \theta) = -y f_\theta(x) + \Phi(f_\theta(x)), \tag{24}$$

where $z = (x, y) \in \mathbb{R}^d \times \{0, 1\}$. Thus we need to check Assumption 3 to apply Theorem 1.

Lemma 5. *Let the loss function be of the form (24) where Φ satisfies the above assumptions. Then for any $\theta \in \mathbb{R}^M$ we have*

$$\tau(K)\|f_\theta - f_{\theta^*}\|^2 \leqslant \mathcal{E}(\theta).$$

Proof. For any $\theta \in \Theta$, we have

$$\begin{aligned} Q(Z, \theta) - Q(Z, \theta^*) &= \nabla Q(Z, \theta^*)^T(\theta - \theta^*) \\ &\quad + \left[\int_0^1 \Phi^{(2)}(H(X)^T(\theta^* + t(\theta - \theta^*)))(1 - t)dt\right] f_\Delta(X)^2 \\ &\geqslant \nabla Q(Z, \theta^*)^T(\theta - \theta^*) + \tau(\|f_\theta\|_\infty \vee \|f_{\theta^*}\|_\infty) f_\Delta(X)^2. \end{aligned}$$

Since $\|\nabla Q(\cdot, \cdot)\|_\infty < \infty$, we can differentiate under the expectation sign, so that

$$\mathbb{E}(\nabla Q(Z, \theta^*)^T(\theta - \theta^*)) = \nabla R(\theta^*) = 0.$$

Thus

$$\mathcal{E}(\theta) \geqslant \tau(\|f_\theta\|_\infty \vee \|f_{\theta^*}\|_\infty)\|f_\theta - f_{\theta^*}\|^2.$$

□

Thus Assumption 3 is satisfied with the constants $\kappa = 2$ and $c = \frac{1}{\sqrt{\tau(K)}}$. We have the following corollary of Theorem 1.

Corollary 2. *Let Assumptions 1, 2 and 4 be satisfied. If $M(\theta^*) \leqslant s$, then, with probability at least $1 - M^{-1} - M^{-K} - 3M^{-2K} \log \frac{n}{\log M}$, we have*

$$\sup_{\hat{\theta} \in \hat{\Theta}} \mathcal{E}(\hat{\theta}) \leqslant \frac{4(1 + 2K)^2}{\tau(K)\zeta(s)^2} s r^2 + \frac{2}{3} r^2,$$

and

$$\sup_{\hat{\theta}\in\hat{\Theta}} |\hat{\theta} - \theta^*|_1 \leqslant \frac{4(1+2K)}{\tau(K)\zeta(s)^2} sr + \frac{K}{3(1+2K)} r.$$

4 Sup-norm convergence rate for the regression model with the Huber loss

In this section, we derive the sup-norm convergence rate of the Dantzig estimators to the target vector θ^* in the linear regression model under a mutual coherence assumption on the dictionary and Huber's loss. The proof relies on the fact that the Hessian matrix of the risk also satisfies the mutual coherence condition for this particular model. Unfortunately, we cannot proceed similarly in the general case because the Hessian matrix of the risk at point θ^* does not necessarily satisfy the mutual coherence condition even if the Gram matrix of the dictionary does. Note that for Huber's loss the Dantzig minimization problem (2) is computable feasible. The constraints in (2) are indeed linear, so that (2) is a linear programming problem.

Denote by $\Psi(\theta)$ the Hessian matrix of the risk R evaluated at θ. With our assumptions on the dictionary $\mathcal{D}$ and on the function γ, for any $\theta \in \mathbb{R}^M$ we have

$$\Psi(\theta) \stackrel{\triangle}{=} \nabla^2 R(\theta) = \left(\mathbb{E}\left(\partial_u^2 \gamma(Y, f_\theta(X)) f_j(X) f_k(X)\right)\right)_{1\leqslant j,k\leqslant M}.$$

Note that for the quadratic loss we have $\Psi(\cdot) \equiv 2G$ where G is the Gram matrix of the design. For Lipschtiz loss functions the Hessian matrix Ψ varies with θ.

We consider the linear regression model (20). For any functions $g, h : \mathcal{X} \to \mathbb{R}$, denote by $< g, h >$ the scalar product $\mathbb{E}(g(X)h(X))$. Define the Gram matrix G by

$$G = (< f_j, f_k >)_{1\leqslant j,k\leqslant M}.$$

From now on, we assume that G satisfies a mutual coherence condition.

Assumption 5. *The Gram matrix $G = (< f_j, f_k >)_{1\leqslant j,k\leqslant M}$ satisfies*

$$G_{j,j} = 1, \forall 1 \leqslant j \leqslant M,$$

and

$$\max_{j\neq k} |G_{j,k}| \leqslant \frac{1}{3\beta s},$$

where $s \geqslant 1$ is an integer and $\beta > 1$ is a constant.

This assumption is stronger than Assumption 1. We have indeed the following Lemma (cf. Lemma 2 in [15]).

Lemma 6. *Let Assumption 5 be satisfied. Then*

$$\zeta(s) \stackrel{\triangle}{=} \min_{J\subset\{1,\cdots,M\},|J|\leqslant s} \min_{\Delta\neq 0:|\Delta_{J^c}|_1\leqslant|\Delta_J|_1} \frac{\|f_\Delta\|}{|\Delta_J|_2} \geqslant \sqrt{1-\frac{1}{\beta}} > 0.$$

Note that Assumption 5 the vector θ^* satisfying (20) such that $M(\theta^*) \leqslant s$ is **unique**. Consider indeed two vectors θ^1 and θ^2 satisfying (20) such that $M(\theta^1) \leqslant s$ and $M(\theta^2) \leqslant s$. Denote $\theta = \theta^1 - \theta^2$ and $J = J(\theta^1) \cup J(\theta^2)$. Clearly we have $f_\theta(X) = 0$ a.s. and $M(\theta) \leqslant 2s$. Assume that θ^1 and θ^2 are distinct. Then,

$$\begin{aligned}\frac{\|f_\theta\|_2^2}{|\theta|_2^2} &= 1 + \frac{\theta^T(G - I_M)\theta}{|\theta|_2^2} \\ &\geqslant 1 - \frac{1}{3\beta s}\sum_{i,j=1}^{M}\frac{|\theta_i||\theta_j|}{|\theta|_2^2} \\ &\geqslant 1 - \frac{1}{3\beta} > 0,\end{aligned}$$

where we have used the Cauchy-Schwarz inequality. This contradicts the fact that $f_\theta(X) = 0$ a.s.

For the linear regression model, the Hessian matrix Ψ at point θ is

$$\Psi(\theta) = \mathbb{E}(\mathbb{1}_{|f_{\theta^*-\theta}(X)+W| \leqslant 2K+\alpha} f_j(X) f_k(X))_{1 \leqslant j,k \leqslant M}.$$

We observe that

$$\Psi(\theta^*) = \mathbb{P}(|W| \leqslant 2K + \alpha)G.$$

Thus $\Psi(\theta^*)$ satisfies a condition similar to Assumption 4 but with a different constant if $\mathbb{P}(|W| \leqslant 2K + \alpha) > 0$. The empirical Hessian matrix $\hat{\Psi}$ at point $\theta \in \mathbb{R}^M$ is defined by

$$\hat{\Psi}_{j,k}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{|f_{\theta^*-\theta}(X_i)+W_i| \leqslant 2K+\alpha} f_j(X_i) f_k(X_i),\ 1 \leqslant j,k \leqslant M.$$

We will prove that the empirical Hessian matrix $\hat{\Psi}(\theta)$ satisfies a mutual coherence condition for any θ in a small neighborhood of θ^* under some additional assumptions given below.

First, we need an additional mild assumption on the noise.

Assumption 6. *There c.d.f. F_W of W is Lipschitz continuous.*

This assumption is satisfied, e.g., if W admits a bounded density so we allow heavy tailed distributions such as the Cauchy. In the sequel we assume w.l.o.g. that the Lipschitz constant of F_W equals 1.

We impose a restriction on the sparsity s.

Assumption 7. *The sparsity s satisfies $s \leqslant \frac{1}{\sqrt{r}}$.*

This implies that we can recover the sparse vectors with at most $O\left((n/\log M)^{1/4}\right)$ nonzero components.

Define $V_\eta = \{\theta \in \Theta : |\theta - \theta^*|_1 \leqslant \eta\}$ where $\eta = C_1 rs$ and

$$C_1 = \frac{8(1+2K)\beta}{\mathbb{P}(|W| \leqslant \alpha)(\beta - 1)} + \frac{1}{6}. \tag{25}$$

Consider the event

$$E = \left\{ \sup_{1\leqslant j,k\leqslant M, \theta\in V_\eta} \left| \hat{\Psi}_{j,k}(\theta) - \Psi_{j,k}(\theta) \right| \leqslant 8L^3\eta + 4L\tilde{r} + \frac{C_2}{\sqrt{n}s^2} \right\}, \tag{26}$$

where

$$C_2 = 2\sqrt{1 + (1+L^2)\left(8C_1L^3 + \frac{4L}{s}\right)} + \frac{1+L^2}{3}.$$

We have the following intermediate result.

Lemma 7. *Let Assumptions 2- 6 be satisfied. Then* $\mathbb{P}(E) \geqslant 1 - \exp(-\sqrt{\log M})$.

Proof. Define the variable

$$Z = \sup_{1\leqslant j,k\leqslant M,\, \theta\in V_\eta} \left| \hat{\Psi}_{j,k}(\theta) - \Psi_{j,k}(\theta) \right|.$$

Applying the Bousquet concentration inequality (cf. Theorem 4 in Section 6) with the constants $c = (1+L^2)/n$, $\sigma^2 = 2/n^2$ and $x = \frac{\sqrt{n}}{s^2}$ yields that, with probability at least $1 - e^{-x}$,

$$Z \quad \leqslant \quad \mathbb{E}\left(Z\right) + \frac{2}{\sqrt{n}s}\sqrt{1 + (1+L^2)\mathbb{E}\left(Z\right)} + \frac{1+L^2}{3\sqrt{n}s^2}. \tag{27}$$

We study now the quantity $\mathbb{E}(Z)$. A standard symmetrization and contraction argument yields

$$\begin{aligned}
\mathbb{E}(Z) &\leqslant 2\mathbb{E}\left(\sup_{1\leqslant j,k\leqslant M,\, \theta\in V_\eta} \left| \frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{1}_{|f_{\theta^*-\theta}(X_i)+W_i|\leqslant 2K+\alpha} f_j(X_i) f_k(X_i) \right| \right) \\
&\leqslant 2\mathbb{E}\left(\left| \frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{1}_{|W_i|\leqslant 2K+\alpha} f_j(X_i) f_k(X_i) \right| \right) \\
&\quad + 2\mathbb{E}\left(\sup_{1\leqslant j,k\leqslant M,\, \theta\in V_\eta} \left| \frac{1}{n}\sum_{i=1}^n \epsilon_i (\mathbb{1}_{|f_{\theta^*-\theta}(X_i)+W_i|\leqslant 2K+\alpha} - \mathbb{1}_{|W_i|\leqslant 2K+\alpha}) f_j(X_i) f_k(X_i) \right| \right).
\end{aligned} \tag{28}$$

Denote by (I) and (II) respectively the first term and the second term on the right hand side of the above expression. The contraction principle yields

$$(I) \leqslant 4\mathbb{E}\left(\max_{1\leqslant j,k\leqslant M} \left| \frac{1}{n}\sum_{i=1}^n \epsilon_i f_j(X_i) f_k(X_i) \right| \right). \tag{29}$$

Then, similarly as in the proof of Lemma 2 we get

$$\mathbb{E}\left(\max_{1\leqslant j,k\leqslant M} \left| \frac{1}{n}\sum_{i=1}^n \epsilon_i f_j(X_i) f_k(X_i) \right| \right) \leqslant L\tilde{r}.$$

Thus, for (II) we have

$$\begin{aligned}(II) &\leqslant 2L^2\mathbb{E}\left(\sup_{\theta\in V_\eta}\frac{1}{n}\sum_{i=1}^{n}|\mathbb{1}_{|f_\Delta(X_i)+W_i|\leqslant 2K+\alpha}-\mathbb{1}_{|W_i|\leqslant 2K+\alpha}|\right)\\ &\leqslant 2L^2\mathbb{P}\left(2K+\alpha-L\eta\leqslant|W|\leqslant 2K+\alpha+L\eta\right)\\ &\leqslant 8L^3\eta. \end{aligned} \tag{30}$$

Assumptions 4 and 7 yield that $s\leqslant\left(\frac{n}{\log M}\right)^{1/4}$. Combining (27)-(30) yields the result. □

We need an additional technical assumption.

Assumption 8. *We have* $12L^3\eta+L\tilde{r}+\frac{C_2}{\sqrt{n}s^2}\leqslant\frac{\mathbb{P}(|W|\leqslant 2K+\alpha)}{2}$.

This is a mild assumption. It is indeed satisfied for n large enough if we assume that $\mathbb{P}(|W|\leqslant 2K+\alpha)>0$ since Assumption 6 implies that $r\to 0$ as $n\to\infty$.

We have the following result on the empirical Hessian matrix.

Lemma 8. *Let Assumptions 2-8 be satisfied. Then, with probability at least* $1-\exp(-\sqrt{\log M})$*, for any* $\theta\in V_\eta$*, we have*

$$\begin{aligned}\min_{1\leqslant j\leqslant M}|\hat{\Psi}_{j,j}(\theta)| &\geqslant \frac{\mathbb{P}(|W|\leqslant 2K+\alpha)}{2},\\ \max_{j\neq k}|\hat{\Psi}_{j,k}(\theta)| &\leqslant \frac{C_3}{s},\end{aligned} \tag{31}$$

where $C_3=\frac{1}{3\beta}+12L^3C_1+\frac{C_2}{\sqrt{n}s}$.

Proof. For any θ in V_η and any j,k in $\{1,\dots,M\}$ we have

$$\Psi_{j,k}(\theta)-\Psi_{j,k}(\theta^*)=\mathbb{E}\left((\mathbb{1}_{|f_\Delta(X)+W|\leqslant 2K+\alpha}-\mathbb{1}_{|W|\leqslant 2K+\alpha})f_j(X)f_k(X)\right),$$

where $\Delta=\theta-\theta^*$. Then

$$\begin{aligned}|\Psi_{j,k}(\theta)-\Psi_{j,k}(\theta^*)| &\leqslant L^2\mathbb{E}\left(|\mathbb{1}_{|f_\Delta(X)+W|\leqslant 2K+\alpha}-\mathbb{1}_{|W|\leqslant 2K+\alpha}|\right)\\ &\leqslant L^2\mathbb{P}\left(|W|\leqslant 2K+\alpha\,,\,|f_\Delta(X)+W|>2K+\alpha\right)\\ &+ L^2\mathbb{P}\left(|W|>2K+\alpha\,,\,|f_\Delta(X)+W|\leqslant 2K+\alpha\right).\end{aligned}$$

Recall that $|f_\Delta(X)|\leqslant L\eta$. Then

$$\begin{aligned}|\Psi_{j,k}(\theta)-\Psi_{j,k}(\theta^*)| &\leqslant L^2\mathbb{P}\left(2K+\alpha-L\eta\leqslant|W|\leqslant 2K+\alpha+L\eta\right)\\ &\leqslant 2L^2\mathbb{P}\left(2K+\alpha-L\eta\leqslant W\leqslant 2K+\alpha+L\eta\right)\\ &\leqslant 4L^3\eta,\end{aligned} \tag{32}$$

where we have used the fact that the distribution of W is symmetric w.r.t. 0 in the second line and Assumption 6 in the last line. Lemma 7 and (32) yield that, on the event E, for any $\theta \in V_\eta$,

$$\min_{1 \leqslant j \leqslant M} \hat{\Psi}_{j,j}(\theta) \geqslant \mathbb{P}(|W| \leqslant 2K + \alpha) - 12L^3\eta - \frac{C_2}{\sqrt{n}s^2},$$

and

$$\max_{j \neq k} |\Psi_{j,k}(\theta)| \leqslant \frac{C_3}{s}.$$

□

Now we can derive the optimal sup-norm convergence rate of the Dantzig estimators.

Theorem 2. *Let Assumptions 2-8 be satisfied. If $M(\theta^*) \leqslant s$, then, on an event of probability at least $1 - M^{-1} - M^{-K} - \exp(-\sqrt{\log M}) - 3M^{-2K}\log\frac{n}{\log M}$, we have*

$$\sup_{\hat{\theta} \in \hat{\Theta}} |\hat{\theta} - \theta^*|_\infty \leqslant C_4 r,$$

where r is defined in Theorem 1,

$$C_4 = \frac{4 + 2C_1C_3}{\mathbb{P}(|W| \leqslant 2K + \alpha)},$$

with C_1 and C_3 defined respectively in (25) and Lemma 8.

Proof. For any $\hat{\theta}$ in $\hat{\Theta}$ we have

$$\nabla R_n(\hat{\theta}) - \nabla R_n(\theta^*) = \left[\int_0^1 \hat{\Psi}(\theta^* + t\Delta)dt\right]\Delta,$$

where $\Delta = \hat{\theta} - \theta^*$.

The definition of our estimator, Lemma 3 and Corollary 1 yield that, on an event $\mathcal{A}_1$ of probability at least $1 - M^{-1} - \exp(-\sqrt{\log M}) - 3M^{-2K}\log\frac{n}{\log M}$, we have that $\hat{\theta} \in V_\eta$ and

$$\left|\left[\int_0^1 \hat{\Psi}(\theta^* + t\Delta)dt\right]\Delta\right|_\infty \leqslant 2r.$$

Lemma 8 yields that, on the event $\mathcal{A}_1 \cap E$,

$$\frac{\mathbb{P}(|W| \leqslant 2K + \alpha)}{2}|\Delta|_\infty \leqslant 2r + \frac{C_3}{s}|\Delta|_1,$$

so that

$$|\Delta|_\infty \leqslant C_4 r.$$

□

Note that Theorem 2 holds true for the Lasso estimators (2) with exactly the same proof, provided that a result similar to Theorem 1 is valid for the Lasso estimators. This is in fact the case (cf. [22, 12]).

5 Sign concentration property

Now we study the sign concentration property of the Dantzig estimators. We need an additional assumption on the magnitude of the nonzero components of θ^*.

Assumption 9. *We have*

$$\rho \stackrel{\Delta}{=} \min_{j \in J(\theta^*)} |\theta_j^*| > 2C_4 r,$$

where r is defined in Theorem 1 and C_4 is defined in Theorem 2.

We can find similar assumptions on ρ in the work on sign consistency of the Lasso estimator mentioned above. More precisely, the lower bound on ρ is of the order $(s(\log M)/n)^{1/4}$ in [17], $n^{-\delta/2}$ with $0 < \delta < 1$ in [25, 27], $\sqrt{(\log Mn)/n}$ in [3], $\sqrt{s(\log M)/n}$ in [26] and r in [15].

We introduce the following thresholded version of our estimator. For any $\hat{\theta} \in \hat{\Theta}$ the associated thresholded estimator $\tilde{\theta} \in \mathbb{R}^M$ is defined by

$$\tilde{\theta}_j = \begin{cases} \hat{\theta}_j, & \text{if } |\hat{\theta}_j| > C_4 r, \\ 0 & \text{elsewhere.} \end{cases}$$

Denote by $\tilde{\Theta}$ the set of all such $\tilde{\theta}$. We have first the following non-asymptotic result that we call sign concentration property.

Theorem 3. *Let Assumptions 2 and 5-9 be satisfied. If $M(\theta^*) \leqslant s$, then*

$$\mathbb{P}\left(\vec{\text{sign}}(\tilde{\theta}) = \vec{\text{sign}}(\theta^*),\ \forall \tilde{\theta} \in \tilde{\Theta}\right) \geqslant 1 - M^{-1} - M^{-K} - \exp(-\sqrt{\log M}) - 3M^{-2K} \log \frac{n}{\log M}.$$

Theorem 3 guarantees that the sign vector of every vector $\tilde{\theta} \in \tilde{\Theta}$ coincides with that of θ^* with probability close to one.

Proof. Theorem 2 yields $\sup_{\hat{\theta} \in \hat{\Theta}} |\hat{\theta} - \theta^*|_\infty \leqslant C_3 r$ on an event $\mathcal{A}$ of probability at least $1 - 6M^{-1}$. Take $\hat{\theta} \in \hat{\Theta}$. For $j \in J(\theta^*)^c$, we have $\theta_j^* = 0$, and $|\hat{\theta}_j| \leqslant c_2 r$ on $\mathcal{A}$. For $j \in J(\theta^*)$, we have $|\theta_j^*| \geqslant 2C_3 r$ and $|\theta_j^*| - |\hat{\theta}_j| \leqslant |\theta_j^* - \hat{\theta} C_j| \leqslant C_3 r$ on $\mathcal{A}$. Since we assume that $\rho > 2C_3$, we have on $\mathcal{A}$ that $|\hat{\theta}_j| \geqslant> c_2 r$. Thus on the event $\mathcal{A}$ we have: $j \in J(\theta^*) \Leftrightarrow |\hat{\theta}_j| > c_2 r$. This yields $\text{sign}(\tilde{\theta}_j) = \text{sign}(\hat{\theta}_j) = \text{sign}(\theta_j^*)$ if $j \in J(\theta^*)$ on the event $\mathcal{A}$. If $j \notin J(\theta^*)$, $\text{sign}(\theta_j^*) = 0$ and $\tilde{\theta}_j = 0$ on $\mathcal{A}$, so that $\text{sign}(\tilde{\theta}_j) = 0$. The same reasoning holds true simultaneously for all $\hat{\theta} \in \hat{\Theta}$ on the event $\mathcal{A}$. Thus, we get the result. □

6 Appendix

We recall here some well-known results of the theory of empirical processes.

Theorem 4 (Bousquet's version of Talagrand's concentration inequality [2]). *Let X_i be independent variables in $\mathcal{X}$ distributed according to P, and $\mathcal{F}$ be a set of functions from $\mathcal{X}$ to $\mathbb{R}$ such that $\mathbb{E}(f(X)) = 0$, $\|f\|_\infty \leqslant c$ and $\|f\|^2 \leqslant \sigma^2$ for any $f \in \mathcal{F}$. Let $Z = \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i)$. Then with probability $1 - e^{-x}$, it holds that*

$$Z \leqslant \mathbb{E}(Z) + \sqrt{2x(n\sigma^2 + 2c\mathbb{E}(Z))} + \frac{cx}{3}.$$

Theorem 5 (Symmetrization theorem [24], p. 108). *Let $X_1, \ldots, X_n$ be independent random variables with values in $\mathcal{X}$, and let $\epsilon_1, \ldots, \epsilon_n$ be a Rademacher sequence independent of $X_1, \ldots, X_n$. Let $\mathcal{F}$ ba a class of real-valued functions on $\mathcal{X}$. Then*

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n (f(X_i) - \mathbb{E}(f(X_i)))\right|\right) \leqslant 2\mathbb{E}\left(\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \epsilon_i f(X_i)\right|\right).$$

Theorem 6 (Contraction theorem [14], p. 95). . *Let $x_1, \ldots, x_n$ be nonrandom elements of $\mathcal{X}$, and let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$. Consider Lipschitz functions $\gamma_i :\to \mathbb{R}$, that is,*

$$|\gamma_i(s) - \gamma_i(s')| \leqslant |s - s'|, \quad \forall s, s' \in \mathbb{R}.$$

Let $\epsilon_1, \ldots, \epsilon_n$ be a Rademacher sequence. Then for any function $f^* : \mathcal{X} \to \mathbb{R}$, we have

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \epsilon_i(\gamma_i(f(x_i)) - \gamma_i(f^*(x_i)))\right|\right) \leqslant 2\mathbb{E}\left(\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \epsilon_i((f(x_i) - f^*(x_i))\right|\right).$$

Acknowledgements: I wish to thank Pr. Alexandre Tsybakov for his useful advices.

References

[1] P.J. Bickel, Y.Ritov and A.B. Tsybakov (2007). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist*, to appear. Preprint available at http://arxiv.org/abs/0801.1095.

[2] O. Bousquet (2002). A Bennet concentration inequality and its application to suprema of empirical processes. *C.R. Math. Acad. Sci. Paris* **334**, 495-550.

[3] F. Bunea (2007). Consistent selection via the Lasso for high-dimensional approximating regression models. *IMS Lecture Notes-Monograph Series*, to appear.

[4] F. Bunea, A.B. Tsybakov and M.H. Wegkamp (2007). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics* **1**, 169-194.

[5] F. Bunea, A.B. Tsybakov and M.H. Wegkamp (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35** 4, 1674-1697.

[6] S.S. Chen, D.L. Donoho and M.A. Saunders (1999). Atomic Decomposition by Basis Pursuit. *SIAM Journal on Scientific Computing* **20**, 33-61.

[7] E. Candes and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2392-2404.

[8] D.L. Donoho, M. Elad and V. Temlyakov (2006). Stable recovery of Sparse Overcomplete representations in the Presence of Noise. *IEEE Trans. on Information Theory* **52**, 6-18.

[9] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani (2004). Least angle regression. *Ann. Statist.* **32**, 402-451.

[10] K. Knight and W. J. Fu (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28**, 1356-1378.

[11] V. Koltchinskii (2007). Sparsity in penalized empirical risk minimization. *Ann. IHP*. Probability and Statistics, to appear.

[12] V. Koltchinskii (2008). Dantzig selector and sparsity oracle inequalities. *Bernoulli*, to appear.

[13] V. Koltchinskii (2008). Oracle inequalities in empirical risk minimization and sparse recovery problems. *Saint-Flour Lectures notes.*

[14] M. Ledoux and M. Talagrand (1991). Probability in Banach spaces: Isoperimetry and Processes. Springer, Berlin.

[15] K. Lounici (2008). Sup norm convergence and sign concentration property of the Lasso and Dantzig estimators. *Electronic Journal of Statistics* **2**, 90-102.

[16] N. Meinshausen and P. Bühlmann (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34**, 1436-1462.

[17] N. Meinshausen and B. Yu (2006). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, to appear.

[18] A. Nemirovski (2000). Topics in nonparametric statistics. In *Lectures on probability theory and statistics (Saint Flour, 1998), Lecture Notes in Math., vol. 1738.* Springer, Berlin, 85 - 277.

[19] M.R. Osborne, B. Presnell and B.A. Turlach (2000a). On the Lasso and its dual. *Journal of Computational and Graphical Statistics* **9** 319-337.

[20] R. Tibshirani (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267-288.

[21] A.B. Tsybakov (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 1, 135-166.

[22] S.A. Van der Geer (2008). High dimensional generalized linear models and the Lasso. *Ann. Statist.* **36** 2, 614-645.

[23] S.A. Van der Geer (2007). The Deterministic Lasso. Tech Report n°140, Seminar für Statistik ETH, Zürich.

[24] A. Van der Vaart and J. Wellner (1996). Weak Convergence and Empirical Processes: With Applications to Statistics. Springer Series in Statistics.

[25] M.J. Wainwright (2006). Sharp thresholds for noisy and high-dimensional recovery of sparsity using l_1-constrained quadratic programming. Technical report 709, Department of Statistics, UC Berkeley.

[26] C.H. Zhang and J. Huang (2008). The sparsity and biais of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **36** 4, 1567-1594.

[27] P. Zhao and B. Yu (2007). On model selection consistency of Lasso. *Journal of Machine Learning Research* **7**, 2541-2567.

[28] H. Zou (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101** n°476, 1418-1429.